

NCBI Mini-Course:

Making Sense of DNA and Protein Sequences

Medha Bhagwat, NCBI

Outline

About NCBI

NCBI Mini-courses

Making Sense of DNA and Protein Sequences

Eukaryotic DNA query

Predict coding region/exons

Obtain protein product

Identify motif/site

Search for similar sequences

Predict function

Perform multiple sequence alignment

Obtain 3-D structural template

Outline

Making Sense of DNA and Protein Sequences
Eukaryotic DNA query (Drosophila genome)
Predict coding region/exons (GenScan)
Obtain protein product (GenScan)
Identify motif/site (ScanProsite)
Search for similar sequences (BLASTp)
Predict function (COG)
Perform multiple sequence alignment (Multalin)
Obtain 3-D structural template (CDD)



hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by [PS00105](#) AA_TRANSFER_CLASS_1 *Aminotransferases class-I pyridoxal-phosphate attachment site* :

[16-51-57-](#)
[GENS~](#)

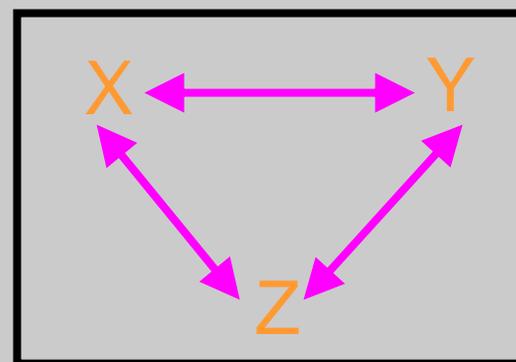


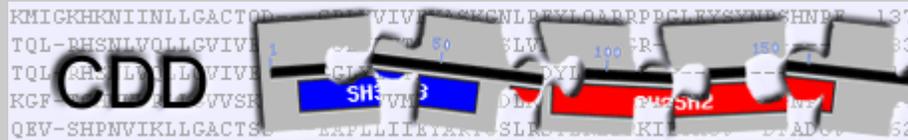
270 - 283: SFAKnmGLyGeRAG



Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes. A COG consists of individual proteins or groups of paralogs from at least 3

Code	Name	Proteins in COG
◇ A	Archaeoglobus fulgidus	2420 1872
◇ O	Halobacterium sp. NRC-1	2605 1701
◇ M	Methanococcus jannaschii	1786 1330
	Methanobacterium thermoautotrophicum	1873 1388
◇ P	Thermoplasma acidophilum	1482 1230
	Thermoplasma volcanium	1499 1243
◇ K	Pyrococcus horikoshii	1800 1378
	Pyrococcus abyssi	1768 1456
◇ Z	Aeropyrum pernix	1841 1178
◇ Y	Saccharomyces cerevisiae	5955 2290
◇ Q	Aquifex aeolicus	1560 1329
◇ V	Thermotoga maritima	1858 1527
◇ D	Deinococcus radiodurans	3187 2226
◇ R	Mycobacterium tuberculosis	3927 2585
	Mycobacterium leprae	1605 1134
◇ L	Lactococcus lactis	2267 1618
	Streptococcus pyogenes	1697 1211
◇ B	Bacillus subtilis	4118 2870
	Bacillus halodurans	4066 2878
◇ C	Synechocystis	3167 2159
	Escherichia coli K12	4275 3414
◇ E	Escherichia coli O157	5315 3662
	Buchnera sp. APS	575 568
◇ F	Pseudomonas aeruginosa	5567 4392
◇ G	Vibrio cholerae	3835 2820
◇ H	Haemophilus influenzae	1714 1542
	Pasteurella multocida	2015 1751

[Help](#)[COGnitor](#)Genome A
All proteinsGenome O
All proteins



<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

Conserved Domain

- recurring unit in molecular evolution, whose extents can be determined by sequence and structure analysis
- performs a particular function
- represented as a multiple local sequence alignment of proteins containing the domain

Eukaryotic DNA query

LOCUS **AEO03584** 337636 bp DNA linear INV 15-MAR-2004
DEFINITION *Drosophila melanogaster* chromosome 2L, section 7 of 83 of the complete sequence.
ACCESSION AEO03584 AEO02638 AEO02637 AEO14134
VERSION AEO03584.4 GI:45444974
KEYWORDS .
SOURCE *Drosophila melanogaster* (fruit fly)

22561 gtteaactgc ctttegtttt aagacgacc actaaaaaca aaagaacagg aaaccgagcg
22621 aaaaacatgt atagagcgtt aattgagtg gtggaagtcg aatgccectt tgtttttaa

22681
22741 [CDS](#) join(329762..330066,330138..330959,331012..331118,
331177..331673)
22801 /locus_tag="CG7082"
22861 /codon_start=1
22921 /product="CG7082-PA"
22981 /protein_id="[AAF51286.1](#)"
23041 /db_xref="GI:7295989"
23101 /db_xref="FLYBASE:[FBgn0031401](#)"
23161 /translation="MLRNTPPFGATPTYKLLLGFGLCSLGGAMLYAYFKTRNDEEEADS
GGQRPASGIRGQTEEQKPQKEVCLKIVVDNEHVPLIMGRGGSNIKLIIEKTLAKIRLR
DKDSGHKFCDISGVPDAVKAARALLIKEIERAPVVVKVELQVPQRLASKINGRGGELLQ
EIRSSSLAKLNIDLNGRNGKAKITII GNQKQVNIARKMLDDQIEEDEELVRSMEVEEQ
RREPRRSPTNSIASSMYSSQTSLSSTQPRDKLMASKGEGKPMEVVYSAVASPTKFWV
QLIGPQSKKLD SMVQEMTSYYSSAENRAKHVLTAPYVQGIVAAVFKFDEKWYRAEIVD
IMPNQYNPKEQVIDLYFVDYGDSEYISPADICELRTDFLTLRFQAVECFLANVKSTIQ
TEPITWPKSSIAKFEELTEVAHWKLIARVVTYKERPRATTAVSAAAKEGTPLPGVEL
FDPADNSELNIADLMITQGFALPLDDSYVRSRSSTPSSNSDSTIEELCVSNPVTPLT
PHSPMSMSIDVDSITQAENEHLAQQLQHLQHLKNGNDIKNINPAKL TATDLENGNNNN
ASTTNGASAH"

Predict coding region/exons

The New GENSCAN Web Server at MIT

Organism: Suboptimal exon cutoff (optional):

Sequence:

Print options:

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

```
catccacaat
 239101 gatggtgacg ttgacatggg ttttgttgcc gtaattgagg cgcttgaagt
ggaagtcctt
 239161 atgcagaaac ttccgatgga tgtttgcactc cagatgctga ttgctaaaac
gaaagctgcc
 239221 aaagctcagg accatggctt ggatgactcc cgaagctccc gctcgaataa
gattgtggca
 239281 gccctgtttc tccaacgtta gcatccacga cgtaaccagg gagttaagct
gctgaagatt
 239341 agacatctcg cgccaaagat tctccgcttg cagagtgtga tacgaatcgt
agcagccttc
```

To have the results mailed to you, enter your email address here (optional):

The New GENSCAN Web Server at MIT

GENSCAN 1.0 Date run: 11-Oct-104 Time: 16:51:59
Sequence 16:51:57 : 5100 bp : 46.29% C+G : Isochore 2 (43 - 51 C+G%)
Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Sngl	+	27	458	432	2	0	48	49	383	0.447	24.68
1.02	PlyA	+	489	494	6							1.05
2.00	Prom	+	830	869	40							-6.86
2.01	Init	+	1002	1069	68	2	2	53	89	83	0.970	3.88
2.02	Intr	+	2549	2708	160	2	1	72	105	284	0.980	28.49
2.03	Intr	+	2771	2872	102	1	0	10	86	251	0.999	17.47
2.04	Intr	+	2935	3183	249	0	0	73	100	586	0.999	55.93
2.05	Term	+	3253	3948	696	0	0	90	49	1324	0.999	122.25
2.06	PlyA	+	4120	4125	6							1.05
3.04	PlyA	-	4162	4157	6							-0.45
3.03	Term	-	4448	4261	188	0	2	37	42	95	0.922	-2.55
3.02	Intr	-	4635	4511	125	2	2	44	90	91	0.949	5.13
3.01	Init	-	5046	4694	353	0	2	66	43	485	0.897	38.43

The New GENSCAN Web Server at MIT

Predicted peptide sequence(s):

```
>16:51:57|GENSCAN_predicted_peptide_1|143_aa  
MPRTLPTTTVFTAVASSARAKSMEKLTVVFLLRMHSAALVVSQPSMATRVNLPVFDPOSLN  
SRAPAKTTSAQAITAYLSIFFHLIELQGKRIGWLFWRWLSPLSASSQRYESTKSGESPKT  
TQSFMRMNGKQLRAATQKKAFFDD
```

```
>16:51:57|GENSCAN_predicted_peptide_2|424_aa  
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMPPDAILGVTEAFKKDTNPKKINLGAGA  
YRDDNTQPFVLPVSVREA EKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN  
VTAQSIGTGALRIGAAFLAKFWQGNREIYIPSPSWG NHVAIFEHAGLPVNR YRYDYDKDT  
CALDFGGLIEDLKKIPEKSI VLLHACA HNPTGVDPTLEQWREISALVKKRNLYPFIDMAY  
QGFATGDIRDAQAVRTFEADGHDFCLAQSF AKNMGLYGERAGAF TVLCSDEEEAARVMS  
QVKILIRGLYSNPPVHGARIAAEILMNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG  
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH  
KVTK
```

```
>16:51:57|GENSCAN_predicted_peptide_3|221_aa  
MSNLQQNLNLSVTSWMLTLEKQGCHNLIRAGASGVIQAMVLSFGSFRFSNQHLECNHPKF  
LHRDFHFRRLNYGNKTHVNVTIIVDDDNKAVINIALDRSDRSYYACDGGCLDEPVLLTON  
RRQFPVKLTEPLTAILYITEDKQHMEELHHAHVKEVVEAPAHEQHLIALHRHGHQLGGL  
PTLFWVSVCAIIIVFHIFLCKLI IKEYCEPSDKLR YRYNKP
```

Outline

About NCBI

NCBI Mini-courses

Making Sense of DNA and Protein Sequences

Eukaryotic DNA query (Drosophila genome)

Predict coding region/exons (GenScan)

Obtain protein product (GenScan)

Identify motif/site (ScanProsite)

Search for similar sequences (BLASTp)

Predict function (COG)

Perform multiple sequence alignment (Multalin)

Obtain 3-D structural template (CDD)

Search for



Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number (s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**), and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below: (leave this box blank to scan PROSITE entry(s) against selected protein databases)

```
>16:51:57|GENSCAN predicted peptide 2|42
4 aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPDAILG
VTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVSVREAEKRVVSRSLDKEYATIIGIPEFY
NKAIELALGKSKRLAAKHN
VTAQSIGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHV
AIFEHAGLPVNRIRYYDKDT
```

General options

- Exclude [motifs with a high probability of occurrence](#)
- Show low level score
- Do not scan profiles [\[User Manual\]](#)

Show only sequences with at least hit(s)

Maximum of matched sequences

- Graphical rich view Simple HTML output
- Plain text output Plain text fasta output
- Retrieve complete sequences

Your e-mail (optional): (will send results by e-mail)



The ScanProsite tool [\[Help\]](#) allows to scan protein sequence(s) (either from user) for the occurrence of patterns, profiles and rules (motifs) stored in the database(s) for hits by specific motif(s) [\[Reference / Download ps_scan\]](#). can be used to generate your own patterns. You may either:



ScanProsite

hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by [PS00105](#) **AA_TRANSFER_CLASS_1** *Aminotransferases class-I pyridoxal-phosphate attachment site* :

[09-08-01-GENS~](#)  (424 aa)

270 - 283: SFAKnmGLyGERAG

Aminotransferases class-I pyridoxal-phosphate attachment site

Description:

Aminotransferases share certain mechanistic features with other pyridoxal-phosphate dependent enzymes, such as the covalent binding of the pyridoxal-phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-I, currently consists of the following enzymes:

- Aspartate aminotransferase (AAT) (EC 2.6.1.1). AAT catalyzes the reversible transfer of the amino group from L-aspartate to 2-oxoglutarate to form oxaloacetate and L-glutamate. In eukaryotes, there are two AAT isozymes: one is located in the mitochondrial matrix, the second is cytoplasmic. In prokaryotes, only one form of AAT is found (gene aspC).
- Tyrosine aminotransferase (EC 2.6.1.5) which catalyzes the first step in tyrosine catabolism by reversibly transferring its amino group to 2-oxoglutarate to form 4-hydroxyphenylpyruvate and L-glutamate.
- Aromatic aminotransferase (EC 2.6.1.57) involved in the synthesis of Phe, Tyr, Asp and Leu (gene tyrB).
- 1-aminocyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis.
- *Pseudomonas denitrificans* cobC, which is involved in cobalamin biosynthesis.
- Yeast hypothetical protein Y11.060w

AA_TRANSFER_CLASS_1, PS00105; Aminotransferases class-I pyridoxal-phosphate attachment site (PATTERN)

Consensus pattern: [GS] - [LIVMFYTAC] - [GSTA] - K - x(2) - [GSALVN] - [LIVMFA] - x - [GNAR] - {V} - R - [LIVMA] - [GA]
K is the pyridoxal-P attachment site

Sequences known to belong to this class detected by the pattern: ALL

Other sequence(s) detected in Swiss-Prot: 1

270 - 283: SFAKnmGLyGeRAG

[Search](#)

```
>16:51:57|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAAALRCKSTWTFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVSVREA EKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQISISGTGALRIGAAFLAKFWQGNREIYIPSPSWG NHVAIFEHAGLPVMNRYRYDKDT
CALDFGGLIEDLKKIPEKSI VLLHACA HNPTGVDP TLEQWREISALVKKRNL YPFIDMAY
```

[Set subsequence](#) From: To:
[Choose database](#)
[Do CD-Search](#)

 Now: **BLAST!** or

Format

 Show [Graphical Overview](#) [Linkout](#) [Sequence Retrieval](#) [NCBI-gi](#) Alignment [format](#)

 Use new formatter [Masking Character](#) [Masking Color](#)

 Number of: [Descriptions](#) [Alignments](#)
[Alignment view](#)

Format for PSI-BLAST	<input type="text" value="Pairwise"/> <input type="text" value="Pairwise with identities"/> <input type="text" value="query-anchored with identities"/> <input type="text" value="query-anchored without identities"/> <input type="text" value="flat query-anchored with identities"/> <input type="text" value="flat query-anchored without identities"/> <input type="text" value="Hit Table"/>
--------------------------------------	--

Limit results by entrez query	<input type="text" value="organisms"/>
---	--



NCBI

Nucleotide

Protein

Translations

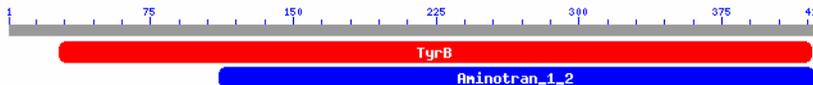
formatting **BLAST**

Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = 09:08:01|GENSCAN_predicted_peptide_2|424_aa(424 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

Format! or **Reset all**

The results are estimated to be ready in 18 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

BLASTP 2.2.14 [May-07-2006]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference:

Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Res.* 29:2994-3005.

RID: 1160659488-27367-133666516875.BLASTQ2

Database: PDB protein database

27,036 sequences; 6,032,163 total letters

gi 112983 sp P00505 AATM_HUMAN	Aspartate aminotransferase, mi...	563	3e-160	
gi 112984 sp P05202 AATM_MOUSE	Aspartate aminotransferase, mi...	562	9e-160	
gi 112987 sp P00507 AATM_RAT	Aspartate aminotransferase, mito...	561	1e-159	
gi 112982 sp P08907 AATM_HORSE	Aspartate aminotransferase, mi...	554	2e-157	
gi 112985 sp P00506 AATM_PIG	Aspartate aminotransferase, mito...	552	9e-157	
gi 1168261 sp P12344 AATM_BOVIN	Aspartate aminotransferase, m...	550	3e-156	
gi 112981 sp P00508 AATM_CHICK	Aspartate aminotransferase, mi...	546	6e-155	
gi 1168256 sp P46643 AAT1_ARATH	Aspartate aminotransferase, m...	447	5e-125	
gi 2506178 sp P28011 AAT1_MEDSA	Aspartate aminotransferase 1 (Tr	434	4e-121	
gi 1168258 sp P46644 AAT3_ARATH	Aspartate aminotransferase, c...	420	4e-117	
gi 21542386 sp P46645 AAT2_ARATH	Aspartate aminotransferase, ...	420	4e-117	
gi 112972 sp P28734 AATC_DAUCA	Aspartate aminotransferase, cytop	420	5e-117	
gi 584706 sp P37833 AATC_ORYSA	Aspartate aminotransferase, cytop	418	2e-116	
gi 112971 sp P00504 AATC_CHICK	Aspartate aminotransferase, cy...	406	8e-113	
gi 21542387 sp P46646 AAT4_ARATH	Aspartate aminotransferase, ...	404	4e-112	
gi 5902703 sp P17174 AATC_HUMAN	Aspartate aminotransferase, c...	402	1e-111	
gi 112976 sp P00503 AATC_PIG	Aspartate aminotransferase, cyto...	397	3e-110	
gi 20532373 sp P46248 AAT5_ARATH	Aspartate aminotransferase, ...	395	1e-109	
gi 112973 sp P08906 AATC_HORSE	Aspartate aminotransferase, cy...	395	2e-109	
gi 112975 sp P05201 AATC_MOUSE	Aspartate aminotransferase, cy...	392	1e-108	

Outline

About NCBI

NCBI Mini-courses

Making Sense of DNA and Protein Sequences

Eukaryotic DNA query (Drosophila genome)

Predict coding region/exons (GenScan)

Obtain protein product (GenScan)

Identify motif/site (ScanProsite)

Search for similar sequences (BLASTp)

Predict function (COG)

Perform multiple sequence alignment (Multalin)

Obtain 3-D structural template (CDD)



COGnitor

Compare your sequence to COG database



compare to COGs

Clear

BeTs to 3 clades

[Help](#)

[Example](#)

Paste your sequence and press the button above.

```
>16:51:57|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAAALRCKSTWFSEVQMGPDDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVREAEKRVSRSRLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSIGTGALRIGAAFLAKFWQGNREIYIPSPSWGPHVAIFEHAGLPVNRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDP TLEQWREISALVKKRNLYPFIDMAY
QGFATGDIRDAQAVRTFEADGHDFCLAQSF AKNMGLYGERAGAF TVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNMEDLRAQWLKDVKLMADRIDVVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHVSVLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
```



16:51:57|GENSCAN_predicted_peptide_2|424_aa (424 letters)

20
proteins

E

[COG1448](#)

Aspartate/aromatic aminotransferase

BeTs to 8 clades

bet-score: 51

[Help](#)

424 letters

904	=>	YLR027c	(432)	-	COG1448
889	=>	NMB0540	(397)	-	COG1448
881	=	NMA0719	(397)	-	COG1448
853	=>	HI1617	(396)	-	COG1448
833	=>	PA3139	(398)	-	COG1448
831	=>	PM0621	(396)	-	COG1448
825	=	aspC	(396)	-	COG1448
819	=	ZaspC	(396)	-	COG1448
788	=>	VC1293	(413)	-	COG1448
781	=	NMB1678	(397)	-	COG1448
780	=	NMA1937	(397)	-	COG1448
724	=>	XF0036	(400)	-	COG1448
710	=	PA0870	(399)	-	COG1448



[20 proteins](#)

[E](#)

[COG1448 info](#)

TyrB

Aspartate/aromatic aminotransferase



[Help](#)

[Genome context](#)

[Pathways /](#) [PHENYLALANINE/TYROSINE BIOSYNTHESIS](#)
[Functional systems](#) [LEUCINE BIOSYNTHESIS](#)

A O M P K Z Y Q V

D R L B C E F G

H S N U J X I T W

A Afu	-
O Hbs	-
M MET	-
P THE	-
K PYR	-
Z Ape	-
Y Sce	YLR027c YKL106w
Q Aae	-

D Dra	-
R MYb	-
L Lla	-
B BAC	-
C Ssp	-
E ENT	tyrB aspC ZaspC ZtyrB
F Pae	PA0870 PA3139
G Vch	VC1293 VCA0513

H PAS	HI1617	PM0621
S Xfa	XF0036	
N Nme	NMB0540 NMB1678	NMA0719 NMA1937
U HPY	-	-
J Mlo	ml10405	-
X Rpr	-	-
I CLA	CT637	CPn0740
T SPI	-	-

>tyrB

MFQKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDEGIIPQLQAVAEAEARLNAOPHGASLYLPME
GLNCYRHAIAPELLFGADHPVLKQQRVATIQTLLGGSGALKVGGADFLKRYFPESGVWVSDPTWENHVAIFAG
AGFEVSTYWPYDEATNGVRFNDLLATLKTLPARSIVLLHPCCHNPTGADLTNDQWDVIEILKARELIPF
LDIAYQGFAGMEEDAYAIRAIASAGLPALVSNSFSKIFSLYGERVGGGLSVMCEDAEAAGRVLGQLKATV
RRNYSSPPNFGAQVVAAVLNDEALKASWLAEEVEEMRTRILAMRQELVKVLSTEMPERNFDYLLNQGMFS
YTGLSAAQVDRLREEFGVYL IASGRMCVAGLNTANVQRVAKAF AAVM

>aspC

MFENITAAPADPILGLADLFRADERPGKINLIGIVYKDETGKTPVLTSVKKAQEYLLNETTKNYLGIDG
IPEFGRCTQELLFGKGSALINDKRARTAQTPGGTGALRVAADFLAKNTSVKRVWVSNPSPWPNHKSVMNSA
GLEVREYAYYDAENHTLDFDALINSLNEAQAGDVVLFHGCCHNPTGIDPTLEQWQTLAQLSVEKGWLP
DFAYQGFARGLEEDAEGRAFAAMHKELIVASSYSKNFGLYNERVGACTLVAADSETVDRAFSQMKAAIR
ANYSNPPAHGASVVATILSNDALRAIWEQELTDMRQRIQMRQLFVNTLQEKGANRDFSFIKQNGMFSF
SGLTKEQVLRRLREEFGVYAVASGRVNVAGMTPDNMAPLCEAIVAVL

>ZaspC

MFENITAAPADPILGLADLFRADERPGKINLIGIVYKDETGKTPVLTSVKKAQEYLLNETTKNYLGIDG
IPEFGRCTQELLFGKGSALINDKRARTAQTPGGTGALRIAADFLAKNTSVKRVWVSNPSPWPNHKSVMNSA
DLEVREYAYYDAENHTLDFDALINSLNEAQAGDVVLFHGCCHNPTGIDPTLEQWQTLAQLSVEKGWLP
DFAYQGFARGLEEDAEGRAFAAMHKELIVASSYSKNFGLYNERVGACTLVAADSETVDRAFSQMKAAIR
ANYSNPPAHGASVVATILSNDALRAIWEQELTDMRQRIQMRQLFVNTLQEKGANRDFSFIKQNGMFSF
SGLTKEQVLRRLREEFGVYAVASGRVNVAGMTPDNMAPLCEAIVAVL

>ZtyrB

MFQKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDEGIIPQLKAVAEAEARLNAVPHGASLYLPME
GLNSYRHAIAPELLFGADHPVLQQRVATIQTLLGGSGALKVGGADFLKRYFPESGVWVSDPTWENHVAIFAG
AGFEVSTYWPYDEATNGVRFNDLLAMLKTLPARSIVLLHPCCHNPTGADLTNDQWDVIEILKARELIPF
LDIAYQGFAGMEEDAYAIRAIASAGLPALVSNSFSKIFSLYGERVGGGLSVLCEDEAEAAGRVLGQLKATV
RRNYSSPPNFGAQVVAAVLNDEALKASWLAEEVEEMRTRILAMRQELVKVLSTEMPERNFDYLLNQGMFS
YTGLSAAQVDRLREEFGVYL IASGRMCVAGLNTANVQRVAKAF AAVM

>PA0870

MSHF AKVARVPGDPILGLLDAYRNDPRADKLDLGVGVYKDAQGLTPILRSVKLAEQRLVEQETTKSYVGG
HGDALFAARLAELALGAASPLLEQRADATQTPGGTGALRLAGDFIAHCLPGRGIWLSDP TWP IHETLFA
AAGLKVSHYPYVSADNRLDVEAMLAGLERIPQGDVLLHACCHNPTGFDLSHDDWRRVLDVVRRELLPL
IDFAYQGFAGGLEEDAMAVRLFAGELPEVLVTSSCSKNFGLYRDRVGALIVCAQNAEKLTDLRSQLAFLA
RNLWSTPPAHGAEVVAAILGDSELKGLWQEEVEGMRSRIASLRIGLVEALAPHGLAERFAHVGAQRGMFS
YTGLSPQQVARLRDEHAVYLVSSGRANVAGLDARRLDRLAQAIQVQCAD

MultiAlin

Multiple sequence alignment by Florence Corpet

Sequence data

Cut and paste your sequences here below.



```
>16:51:57|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVREAEKRVVSRSLDKEYATIIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSIISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNIHVAIFEHAGLPVNRYYDKDT
CALDFGGLIEDLKKIPEKSI VLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIRDAQAVRTFEADGHDFCLAQSF AKNMGLYGERAGFTVLCSDDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNE DLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
|>tyrB
```

(sample sequences)

or select a file: Browse... new

Sequence input format:

For nucleotidic sequences, you must change the Symbol comparison Table (see below)



16:51:57 IGENSCAN_pre
 YLR027c
 tyrB
 ZtyrB
 NMB1678
 NMA1937
 PA3139
 XF0036
 aspC
 ZaspC
 VC1293
 HI1617
 PM0621
 NMB0540
 NMA0719
 VCA0513
 n110405
 PA0870
 CT637
 CPn0740
 YKL106w
 Consensus

MSQICKRGLLISNRLAPAALRCKSTWFSEVQNGPPDAILGVTEAFKKDTPNPKINL GAGAYRDNTPQPFVLPVSVREAEKRYVS-RSLDKE--YATITIGIP
 MNKRTQEYKNTRAIMS----ATL FNNIELLPDALFGIKQRYGQQQRATKVDLGIGAYRDONGKPHVLPVSKAEKLIHNDSSYNHE--YLGITIGLP
 MFQKYDAYAGDPILTLMERFKEDPRSDKYNLSIGLYNEDGIIPQLQAVAEAEARLNAQPHGASL--YLPMEGLN
 MFQKYDAYAGDPILTLMERFKEDPRSDKYNLSIGLYNEDGIIPQLKAVAEAEARLNAVPHGASL--YLPMEGLN
 MYRHIEYYPGDPILSLVETFKNDPRPEKYNLSIGIYFDDEGKMPVLESVSRAEATARAAP-APSP--YLPMEGLD
 MYRHIEYYPGDPILSLVETFKNDPRPEKYNLSIGIYFDDEGKMPVLESVSRAEATARAAP-APSP--YLPMEGLD
 MSLFSAVENAPRDPILGLNEAFNADTRPGKINLGVGYVYNEEGRIPLLRAVQAAEKARIEAH-APRG--YLPTEGIA
 MPLFTDVELVPGDPILSLNDTYNADTRTNKYNL GIGIYCDSESGCIPLLRAVQQVEEQAKHP-KPRG--YLPIDGLP
 MFENITAAPADPILGLADLFRADERPGKINL GIGVYKDETGKTPVLT SVKKAQYLLENE-TTKN--YLGIDGIP
 MFENITAAPADPILGLADLFRADERPGKINL GIGVYKDETGKTPVLT SVKKAQYLLENE-TTKN--YLGIDGIP
 MMSSIIPSHPNLTWCFMFEKVVAAPADPILGLTEEFKKDPRD KINLGVGIYKNEAGETPVLATVYKKAERALLESE-KTKS--YLTIEGTA
 MFEHIKAAPADPILGLGEAFKSETRENKINL GIGVYKDAQGTTPIHNAVKEAEKRLFDE-KTKN--YLTIDGIA
 MFENITAAPADPILGLGESFKAE TRDNKINL GIGVYKDAKGNTPIHNAVKEAEKRLFDLE-HSKN--YLTADGVA
 MFFKHIEAAPADPILGLGEAFKAE TRPEKYNL GIGVYKDA SGATPLVKA VKEAEKRLLESE-TTKN--YLTIDGVA
 MFFKHIEAAPADPILGLGEAFKAE TRPEKYNL GIGVYKDA SGATPIVKA V if the alignment is not visible, try to c
 MLPEFQRV--VTFMHTL PAPVLPDILSLSVAFRNDPRPQKVDLGIGVYKNSLGETPIHNAV and/or text size (see options)
 MFEDLQAPADKILALIGLYRADPRPNKVDLGVGYKORDGKTPVHNAVREAEKRLNSQ-DTKT--YLCGHDG
 MSHFAKVPYVPGDPILGLLDAYRNDPRADKLDLGVGYKDAQGLTILRSVKLAERLVEQE-TTKS--YVGGHGDA
 MSLFEQLPSFSPDILGLAQAFQEDPREDKINLLGT YEREKKRYGGFSSVRKAQSVFFDDE-KDKN--YLPKIGSS
 MSFFNHIPTFSPDAILGLQNVYFADKRPEKYNL VIGVYEHQPKRYGGLSCIRKAQTVILEEE-QNKS--YLPISGLQ
 MLRTRLTNCSLMRPYTSSLSRVPRAPPDKVLGLSEHFKKYKNV NKIDLVGIYKDGAGKVTTFPSVAKAQLIESHLELNKNLSYLPITIGSK

f p De l e % d r K ! # l e i g . Y . d . e . p . l . s v . A # k Y l . i . G

:51:57 IGENSCAN_pre
 YLR027c
 tyrB
 ZtyrB
 NMB1678
 NMA1937
 PA3139
 XF0036
 aspC
 ZaspC
 VC1293
 HI1617
 PM0621
 NMB0540
 NMA0719
 VCA0513
 n110405
 PA0870
 CT637
 CPn0740
 YKL106w
 Consensus

ATGDIIDRDQAVRTE----ADGHDFCLAQSFAK NMGLYGERAGAFV LSCSDE-----EE----AARVMSQVKILIRGLYSNPVYHGARIAEILNNI
 ATGDLKDAYAVRLGV---EKLSTVSPVFCQSFAK NAGMYGERVGC FHLALTKQ-----AQNKTIKPAVTSQ LAKIRSEVSNPPAYGAKIVAKLETT
 GAG-MEEDAYAIRAIA---S--AGLP-ALVSNFSF KIFSLYGERVGGLSV MCEDA-----EA----AGRYLGQLKATVRRNYSPPNFQAQVVAAVLNDI
 GAG-MEEDAYAIRAIA---S--AGLP-ALVSNFSF KIFSLYGERVGGLSV LCEDA-----EA----AGRYLGQLKATVRRNYSPPNFQAQVVAAVLNDI
 GGD-LDSDAYAVRKAV---E--MELP-LFVSNFSF KNL SLYGERVGGLSV VCPNK-----FF----ADLVEGOLKFTVVRTYSSPPAHGAYTAADVHNSI
 GGD-LDSDAYAVRKAV---E--MELP-LFVSNFSF KNL SLYGERVGGLSV VCPNK-----FF----ADLVEGOLKFTVVRTYSSPPAHGAYTAADVHNSI
 GNG-IEEDAAAVRLF A---Q--SGLS-FFVSSSFS KFSLYGERVGGLSV VTESR-----if the alignment is not visible, try to decrease line length S
 and/or text size (see options) S
 NQG-IDADAYAIRLLA---E--EGISNYVVA NSYS KFSLYGERVGGLSV VASNT-----EQ----AQAIQSQVKRIIRTIYSSPSAHGAYLVAGVVLNSI
 ARG-LEEDAREGLRAFA---A--M-HKELIVASSYS KNFGLYNERVGAFTI VAADS-----ET----VDRAFSQMKARIRANYSNP P AHGASVYATILSNI
 ARG-LEEDAREGLRAFA---A--M-HKELIVASSYS KNFGLYNERVGAFTI VAADS-----ET----VDRAFSQMKARIRANYSNP P AHGASVYATILSNI
 ASG-VEEDAGLRIFA---K--Y-NSEILVASSYS KNFGLYNERVGAFTI VAPST-----TY----AETAFSQVKAIRIRSIYSNP P AHGAYVYTYILNNI
 ANG-LEEDAYGLRAFA---A--N-HKELLVASSYS KNFGLYNERVGAFTI VAENA-----EI----ASTSLTQVKSIIIRTLYSNP P AHGATVATVNLNDI
 ANG-LEEDAFGLRTFA---K--N-HKELLVASSYS KNFGLYSERVGAFTI VAETE-----QT----AATALTQVKTIIRTLYSNP P AHGATTVSMVLKDI
 GNG-LEEDAYGLRVFL---K--H-NTELLIASSYS KNFGMYNERVGAFTI VAEDE-----ET----AARAHSQVKTIIRTLYSNP P AHGANTIALVLKNI
 GNG-LEEDAYGLRVFL---K--H-NTELLIASSYS KNFGMYNERVGAFTI VAEDE-----AT----AARAHSQVKTIIRTLYSNP P AHGANTIALVLKNI
 GDG-LEQDAQGLRYMA---E--R-MEENLITTS CS KNFGLYRERTGARTI IGKNQ-----QE----VTNARGKMLTLARSTYTHPPDHGAALVKTVLRDI
 GDG-LEADALGLRLLA---A--K-YPEMVVASSCS KNFAYYRDRVGAAMV LARDS-----AQ----ADVAMSQMLSAARAYSNP P DHGAARVRYVLEDI
 GDG-LEEDAAVRLFA---G--E-LPEVLVTSSCS KNFGLYRDRVGAALV CAQNA-----EK----LTDLRSQAF LARNLWSTPPAHGAEVVAAILGDI
 ASG-IEEDRRPVRLCI---E--AGYTTFVAGGAS KIFSLYGSRVGFFGAIHQDK-----LD----LNRILSFL EEQIRGEYSSPAREGVAVYVTSILSNI
 AHG-IELDRKPIEIFI---S--EGNTVLVASSS KNFALYGERVGYFAYHSTFT-----DE----LVKIH SFL EEKIRGEYSSPQRHGV EIVYSTILSNI
 ESNLLKDAYLLRLCLNVNKYPNMSNGIFLCQSFAK NMGLYGERVGGLSV ITPATANNGKFNPLQKNSLQQNIDSQLKKIVRGHYSPPPGYGSRVVNVVLSDI

..G.le.Da...Rl.a.....va.S.sknf#Yg#RvGa.v.....sqlk.iR..yS.Pp.hGa.v..!L..

Outline

About NCBI

NCBI Mini-courses

Making Sense of DNA and Protein Sequences

Eukaryotic DNA query (Drosophila genome)

Predict coding region/exons (GenScan)

Obtain protein product (GenScan)

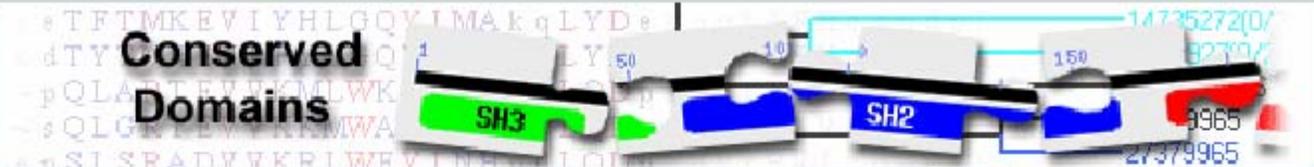
Identify motif/site (ScanProsite)

Search for similar sequences (BLASTp)

Predict function (COG)

Perform multiple sequence alignment (Multalin)

Obtain 3-D structural template (CDD)



Search Conserved Domains on a protein

Search against database:

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format

```

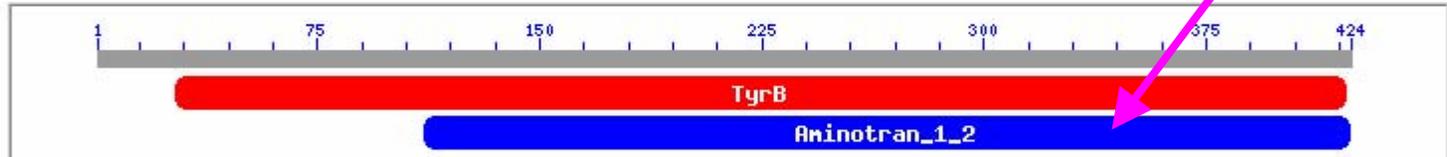
>09:08:01|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPVSVREAERKRVVSRSLDKEYATIIGIPEFYNKAIELALGKSKRLAAKHN
VTAQSIISGTGALRIGAAFLAKFWQGNREIYIPSPSWGHNVAIFEHAGLPVNRYYDKDT
CALDFGGLIEDLKKIPEKSI VLLHACAHNPTGVDPTLEQWREISALVKKRNL YPFIDMAY
QGFATGDIRDAQAVRTFEADGHDFCLAQSF AKNMGLYGERAGFTVLCSEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNE DLRAQMLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDH SVYL TNDGRVSMAGVTSKNVEYLAESI H
    
```

Force live search



Query sequence: [(local sequence)|c| 1]
 09:08:01|GENSCAN_predicted_peptide_2|424_aa

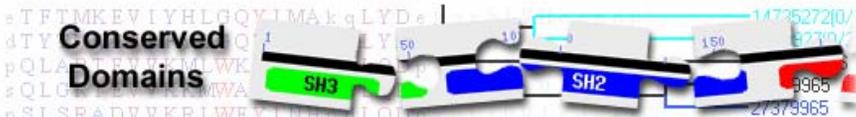
Concise Result Full Result Show Search Information [?](#)



Descriptions

	Title	PssmId	Multi-Dom	E-value
[+]	COG1448, TyrB, Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and ...	31637	No	2e-147
[+]	pfam00155, Aminotran_1_2, Aminotransferase class I and II..	40255	Yes	4e-67

[Search for similar domain architectures](#)



pfam00155.12

Aminotran_1_2, with user query added

Aminotransferase class I and II.

- [+] Links:
- [+] Statistics:
- [-] Structure:

Show Structure

Program:

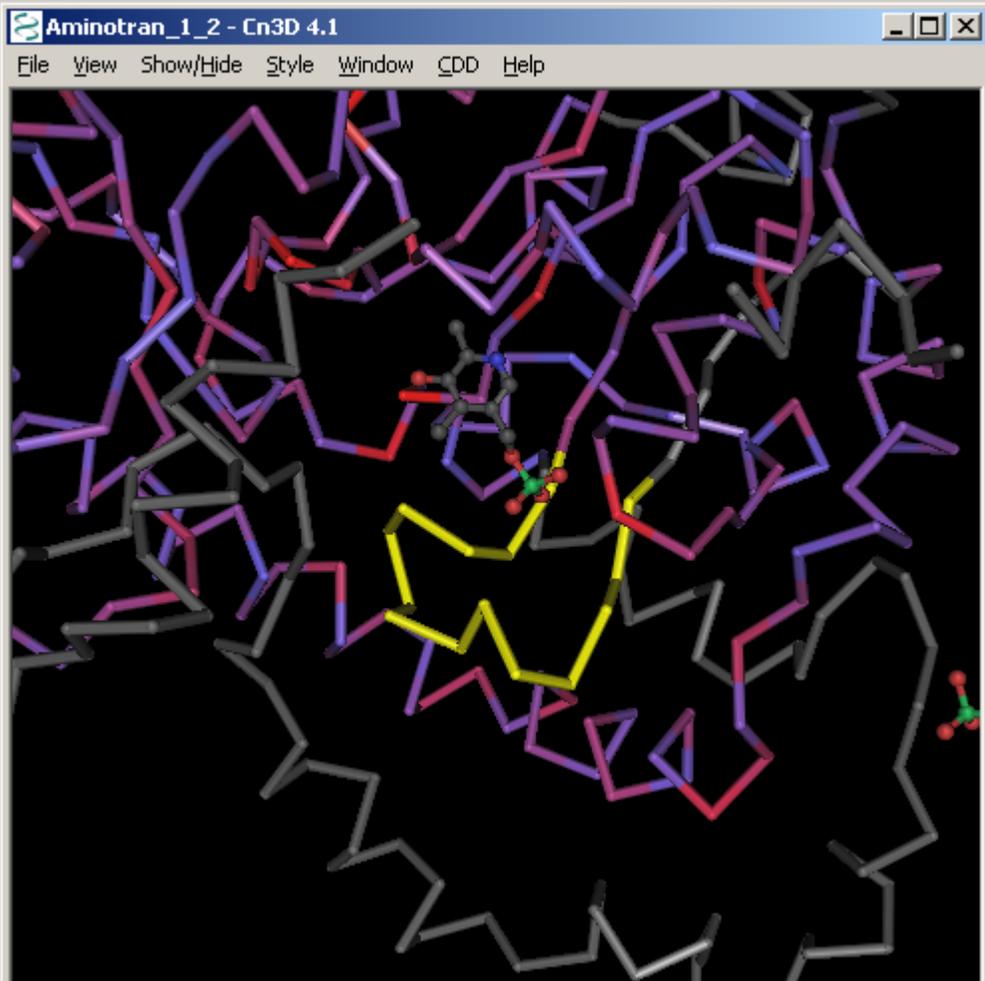
Drawing:

Aligned Rows:

[\[Download Cn3D\]](#)

[\[?\] Reformat Sequence Alignment](#) Format: Row Display: Color Bits: Type Selection:

1B8G_A	95	[3]	.AEIRG	NKVTFDPNHLVLTAGATSANETFIFCLA.	[4]	.AVLIPTPPYPGFDRDLKWRIGV	EIVPI	161
query	108	[3]	.GSKRL	AAKHNVTAQSIGTGALRIGAAFLAKFW.	[4]	.EIYIPSPSWGNNHVAIFEHAGLP.	[1]	.NRYRY 175
1IX6_A	80	[3]	.FGKGS.	[3].NDKRARTAQTGGTGALRVAADFLAKNT.	[3]	.RVWVSNPSWPNHKSVFNSAGLE.	[1]	.REYAY 149
1AJS_A	85	[3]	.LGDDS.	[3].QEKRVGGVQSLGGTGALRIGAEFLARWY.	[8]	.PVYVSSPTWENHNGVFTTAGFK.	[2]	.RSYRY 160
1AMA	82	[3]	.LGENS.	[3].KSGRYVTVQGISGTGSLRVGANFLQRF.	[4]	.DVYLPKPSWGNHTPIFRDAGLQ.	[1]	.QAYRY 152
gi 398985	101	[3]	.FKESC.	[8].AHDRISFVQTLSGTGALAVAAKFLALFI.	[2]	.DIWIPDPSWANHKNIFQNGFPE.	[2]	.YRYSY 175
gi 1168262	80	[3]	.FGKDS.	[3].QSNRARTVQSLGGTGALRIAAEFIKRQT.	[3]	.NVWISTPTWPNHNAIFNAVGMT.	[1]	.REYRY 149
gi 2506178	97	[3]	.FGADS.	[3].QENRVTTVQGLSGTGLRVGGEFLAKHY.	[3]	.IIYLPPTWGNHTKVFNLAGLT.	[1]	.KTYRY 166
gi 21542387	82	[3]	.LGDDS.	[3].KENRVVTTQCLSGTGLRVGAEFLATHN.	[3]	.VIFVVPNTWGNHPRIFTLAGLS.	[1]	.QYFRY 151
gi 1168256	110	[3]	.YGDNS.	[3].KDKRIAAVQTLSGTGACRLFADFQKRFS.	[3]	.QIYIPVPTWNNHNIWKDAQVP.	[1]	.KTYHY 179



CDD Descriptive Items

Name: Aminotran_1_2

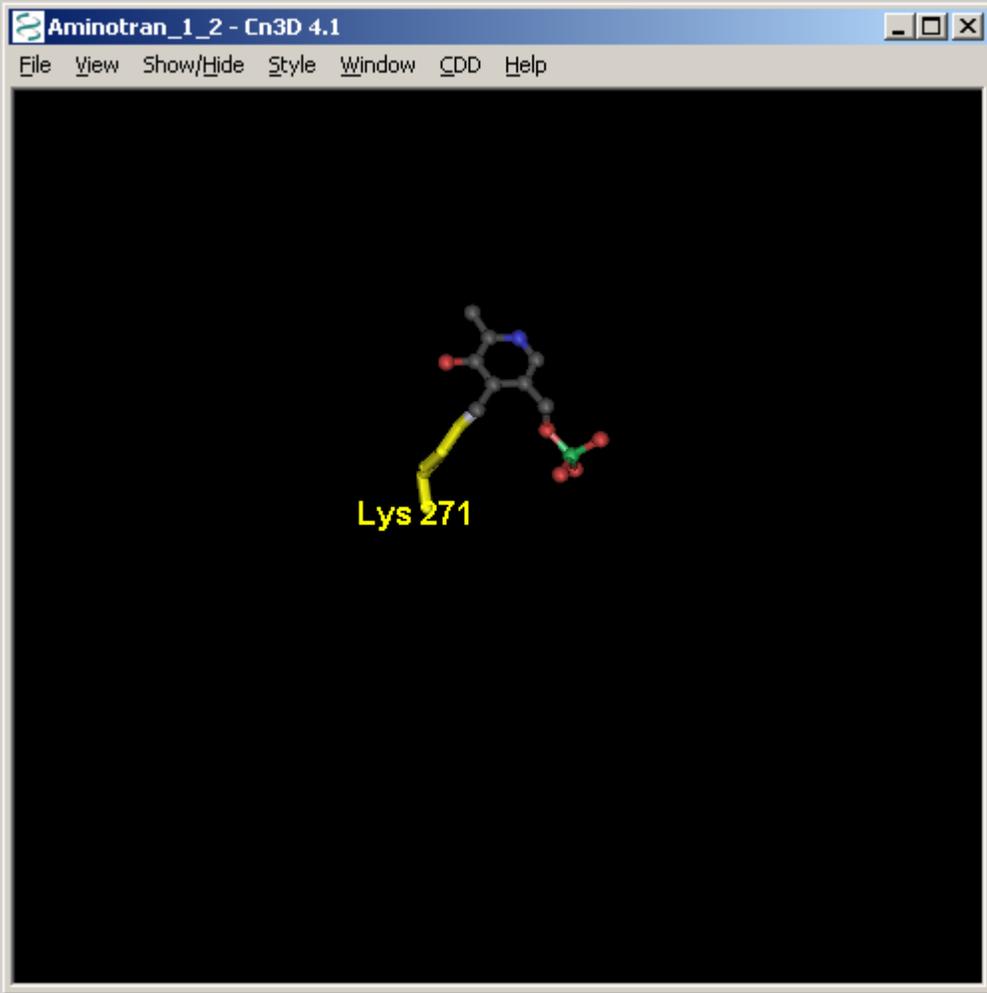
PDB 1B8G (MMDB 12342)
 1B8G_A: gi 6980404 ([Malus x domestica] Chain A,
 1-Aminocyclopropane-1-Carboxylate Synthase)
 1B8G_B: gi 6980405
 Heterogens: PLP (x2)

Show Annotations Panel Show References Panel Dismiss

Aminotran_1_2 - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

<i>1B8G_A</i>	R n c d e n s e v w Q R V H V V Y S L S K D L G L P G F R V G a i y s n d d M V V A A T K M S S F G L V S S Q T Q H L L S A M L S D K K L T K N Y I A E N H K
<i>consensus</i>	Y ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ D N L L V V Q S L S K N F G L A G K R L G g ~ ~ ~ ~ ~ a A G G I V A G S A A S F D R V S S Q S R A L L F A T S S A P P A V G A A I V A L I L
<i>query</i>	G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ H D F C L A Q S F A K N M G L Y G E R A G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A F T V L C S D E E E A A R V M S Q V K I L I R G L Y S N P P V H G A R I A A E I L
<i>IAMA</i>	G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ I D V V L S Q S Y A K N M G L Y G E R A G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A F T V I C R D A E E A K R V E S Q L K I L I R P M Y S N P P M N G A R I A S L I L
<i>gi 2506178</i>	G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ G E L L V A Q S Y A K N M G L Y G E R V G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A L S I V S K S A D V S S R V E S Q L K L V I R P M Y S S P P I H G A S I V A A I L



Aminotran_1_2 - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

<p><i>JB8G_A</i></p> <p>consensus</p> <p>query</p> <p>iAMA</p> <p>gi 2506178</p>	<pre> R n c d e n s e v w Q R V H V V Y S L S DLGLPGFRVG a i y s n d d M V V A A T K M S S F G L V S S Q T Q H L L S A M L S D K K L T K N Y I A E N H K Y ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ D N L L V V Q S L S K N F G L A G K R L G g ~ ~ ~ ~ ~ a A G G I V A G S A A S F D R V S S Q S R A L L F A T S S A P P A V G A A I V A L I L G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ H D F C L A Q S F A K N M G L Y G E R A G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A F T V L C S D E E E A A R V M S Q V K I L I R G L Y S N P P V H G A R I A A E I L G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ I D V V L S Q S Y A K N M G L Y G E R A G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A F T V I C R D A E E A K R V E S Q L K I L I R P M Y S N P P M N G A R I A S L I L G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ G E L L V A Q S Y A K N M G L Y G E R V G ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ A L S I V S K S A D V S S R V E S Q L K L V I R P M Y S S P P I H G A S I V A A I L </pre>
--	--



Outline

About NCBI

NCBI Mini-courses

Making Sense of DNA and Protein Sequences

Eukaryotic DNA query (Drosophila genome)

Predict coding region/exons (GenScan)

Obtain protein product (GenScan)

Identify motif/site (ScanProsite)

Search for similar sequences (BLASTp)

Predict function (COG)

Perform multiple sequence alignment (Multalin)

Obtain 3-D structural template (CDD)

Hands-on Session

<http://www.ncbi.nlm.nih.gov/Class/minicourses/x1a.html>